Supplementary Information, Shaffer et al.: inferring interaction network of resistance markers in cancer cells

We use the recently proposed phixer algorithm [1] on single-cell mRNA expression data to uncover interaction networks between resistance markers in undrugged cancer cells. Our results provide novel insights into interactions leading to coordinated rare-cell expression of resistance markers. We begin by providing an overview of the algorithm and its implementation on mRNA FISH data.

Problem formulation

Consider a set of n resistant marker genes whose levels in an individual cell are given by random variables X_1, X_2, ..., X_n. These random variables take values in the positive integers and represent the mRNA counts of corresponding genes as measured by RNA FISH. Ignoring self edges, a network of n genes contains $n(n-1)$ edges that are directed from one gene to another. An edge from gene i to k signifies a causal effect of X_i on X_k that is mediated directly or through unknown factors not measured in the study. Note that a gene can effect its own, or other gene's expression indirectly through intermediate states, creating feedforward and feedback loops in the network.

Most approaches for uncovering gene interactions rely on mutual information or Bayesian methods that impose restrictions on the inferred network. For instance, Bayesian network approaches assume that there are no cycles (or feedbacks) between genes [2]. Moreover, techniques based on mutual information assume interactions to be undirected, and hence provide no information about causality [2]. Recently, a phixer algorithm was introduced that uses joint observations of random variables X_1, X_2, ..., X_n across samples (cells) to infer directed edges between genes, and also allows the network

to contain feedbacks between genes [1]. We briefly discuss this algorithm below and refer interested readers to [1] for further details.

The phixer algorithm

This algorithm relies on computing ϕ-mixing coefficients, which can be interpreted as strengths of the directed edges in the network. More specifically, the ϕ-mixing coefficient corresponding to the edge from gene k to i is given by

$$\phi\left(X_i|X_k\right) = \max_{q,p\in\{0,1,\dots\}} \left|P\left\{X_i = q|X_k = p\right\} - P\left\{X_i = q\right\}\right|, \; i,k \in \{1,\dots,n\}, \; i \neq k, \tag{1}$$

where $P\left\{X_i = q|X_k = p\right\}$ represents the conditional probability of observing q transcripts for gene i, given that there are p transcripts for gene k. In essence, $\phi\left(X_i|X_k\right)$ is the absolute value of the difference between the conditional and unconditional probability, maximized over all possible values of X_i and X_k.

The ϕ-mixing coefficient computed in (1) has some useful properties, such as, it is always bounded $\phi\left(X_i|X_k\right) \in [0,1]$, and $\phi\left(X_i|X_k\right) = 0$ iff X_i and X_k are independent (i.e., no connection between genes). Unlike mutual information and Pearson's correlation coefficient, the ϕ-mixing coefficient is asymmetric $\phi\left(X_i|X_k\right) \neq \phi\left(X_k|X_i\right)$ and provides information on the direction of influence. Next, we describe another important property that plays a key role in "pruning" edges in the network. It involves checking the following inequality

$$\phi\left(X_i|X_k\right) \leq \min\left\{\phi\left(X_i|X_l\right), \phi\left(X_l|X_k\right)\right\}, \tag{2}$$

for a given gene triplet i, k, and l. If (2) holds, then X_i and X_k are conditionally independent given X_l, and hence the edge from gene k to i can be removed [1].

The above properties lead to the following phixer algorithm:

1. Start with a network of n genes and $n(n-1)$ directed edges. Compute $\phi\left(X_i|X_k\right)$ using (1) for each edge.

2. For each gene triplet i, k, j, verify inequality (2). If it holds, then remove edge from k to i. This is referred to as the pruning step of the algorithm.

3. The edges remaining after checking (2) for all possible gene triplets represents the inferred network. To account for sampling errors, edges with ϕ-mixing coefficients below a selected threshold are also removed.

It turns out that the inferred network is invariant of the order in which the triplets are checked in the pruning step, and is robust to any monotone transformation of the data [1]. Note that the inferred edges have direction, but do not have a sign indicating negative or positive interaction. In principle, once the final network is obtained, a sign can be assigned to an edge based on positive/negative correlation between the corresponding pair of genes.

Application to the cancer data set

We use the phixer algorithm to infer gene interactions in melanoma undrugged cancer cells. RNA FISH was used to measure the expression of housekeeping genes (VGF, CCNA2, GAPDH), melanocyte-specific genes (SOX10,MITF) and resistance markers (VEGFC, AXL, JUN, WNT5A, NGFR, SERPINE1, FGFR1, LOXL2, EGFR, NRG1, PDGFRB, RUNX2, FOSL1, and PDGFC). These measurements provide the number of mRNA transcripts for each of the 19 genes in single cells with a sample size of $\approx 10^4$ cells. Note that $n = 19$ genes lead to $19 \times 18 = 342$ directed edges in the network.

To study rare-cell expression of genes, mRNA counts were converted into Bernoulli random variables, where a gene is either OFF (0) or ON (1) depending on its level being below or above a threshold, respectively. This transformation of data into binary values facilitates efficient computation of the ϕ-mixing coefficients. For a given pair of genes, the probability difference in (1) is calculated for the four different cases of genes being ON/OFF, and ϕ is the maximum value among them. Using this approach, the ϕ-mixing coefficients were computed for each of the 342 edges. Interestingly, our results show that the pruning step of the algorithm removes $\approx 80\%$ of the edges. Removal of

these edges suggests that coordinated rare-cell expression of many resistance marker pairs is simply a result of common upstream regulators. Histogram of the ϕ-mixing coefficient for all remaining edges after the pruning step are shown in Supplementary Fig. 19a, and the corresponding inferred network in Supplementary Fig. 19b. For the sake of visual inspection, we only show the 34 strongest edges in the network. This corresponds to showing edges with $\phi \geq 0.18$.

Insights from the inferred networks

The inferred network of resistance markers for undrugged cancer cells is shown in Supplementary Fig. 19b. Here each edge quantifies the effect of an upstream gene on the probability of a downstream gene being ON. The network highlights the intricate interactions that lead to coordinated rare-cell expression of resistance markers and provides insights into causal pathways. For example, data shows high odd of rare-cell expression for NRG1 with many other resistance markers, such as, VEGFC, AXL, JUN, WNT5A and LOXL2. Our results show that NRG1 is a direct upstream regulator of all these genes (Supplementary Fig. 19b). In many cases the effect is mediated through multiple pathways creating feedforward loops. For example, NRG1 directly effects LOXL2, and also effects it indirectly through VEGFC.

The network also identifies resistance markers that have no incoming edges, but many outgoing edges. Example of such markers can be seen in NRG1 and RUNX2. Results show that all edges coming into EGRF have very small ϕ-mixing coefficients ($\phi < 0.05$) that are not significant. However, we find some weak but significant edges from EGFR to WNT5A, and JUN ($\phi \approx 0.15$). These weak edges were not shown in the network as they are below the cutoff of 0.18. A similar result is seen for PDGFRB, where there are no incoming edges but multiple weak outgoing edges to other genes. In summary, these results suggest that a few key markers (NRG1, RUNX2, EGFR, PDGFRB) maybe the first upstream regulators that drive much of the dependent rare-cell expression seen in the data.

Algorithm implementation

We implement the phixer algorithm for binary data in a m-file script called *binPhix* using MatLab®. The script requires as input a text file which includes binary single-cell gene expression measurements. The first row of this file should be a list of n strings (separated by commas) representing the names of the genes being measured. Each additional row contains the expression level of these n genes (separated by commas) within an individual cell. This expression level maybe 0 (unexpressed gene) or 1 (expressed gene). Additionally, a threshold (*thr* parameter in the script) value between 0 and 1 is required to perform step 3 of the phixer algorithm.

binPhix computes the three steps described in the phixer algorithm section. The output of the script is a $n \times n$ matrix representing the ϕ matrix. Each entry corresponds to a $\phi(X_i|X_k)$ value, which represents the weight of the influence of gene in column k on gene in row i. To construct the inferred network, *binPhix* requires the biograph function. This function uses the ϕ matrix and the list of gene names to create a directed graph with n vertices, each one representing a gene. Biograph also creates an edge for each non zero $\phi(X_i|X_k)$ value in the ϕ matrix. The edge represents the influence of gene k (kth column) on the gene in row i (ith row).

Please be aware that *binPhix* will be unable to construct the graph if the bioinformatics toolbox in not present in the MatLab® version used. When the bioinformatics toolbox is missing, *binPhix* will not display the graph. Alternative, the script will generate a text file called *net.txt*. This file includes the list of all edges (and their respective $\phi(X_i|X_k)$ value) in the inferred network. Each line in the file corresponds to a single edge. Each edge is a string in the form "gene k -> gene i", meaning that the gene k influences gene i. This string will be followed by its correspondent $\phi(X_i|X_k)$ value.

Results shown were obtained using a computer with 8 GB RAM and four 3.4 GHz cores. We include a folder with the *binPhix* script and a text file. The text file contains the single-cell measurements of the undrugged melanoma cancer cells used in the inference of the network shown in previous sections.

References

[1] Nitin Kumar Singh, M. Eren Ahsen, Shiva Mankala, Hyun-Seok Kim, Michael A. White, and M. Vidyasagar. Reverse Engineering Gene Interaction Networks Using the Phi-Mixing Coefficient. *arXiv:1208.4066 [q-bio, stat]*, 2012.

[2] Y. X. Rachel Wang and Haiyan Huang. Review on statistical methods for gene network reconstruction using expression data. *Journal of Theoretical Biology*, 362:53–61, 2014.